УДК 004.8

МЕТОДИКА ПРОВЕРКИ НАБОРА ДАННЫХ, ИСПОЛЬЗУЕМЫХ В МАШИННОМ ОБУЧЕНИИ

Лопатин Г.А. [1], Чусов Г.А. [1], Менисов А.Б.[1]
Научный руководитель – доктор технических наук Менисов А.Б.[1]
[1]Военно-космическая академия им. А.Ф.Можайского
vka@mil.ru

Введение

Стремительное развитие технологий искусственного интеллекта и широкое внедрение больших языковых моделей повышают требования к качеству обучающих наборов данных. Вместе с тем, существующее множество открытых данных часто содержит скрытые смещения и шум, что требует разработки унифицированных методик валидации перед интеграцией в цикл машинного обучения [1].

Основная часть

В последние годы внимание исследователей сместилось от атак на инференс модели к атакам на данные и процесс обучения. Некорректные, смещённые или намеренно отравленные данные могут привести к деградации качества, дискриминации, ложным решениям и др. ущербу. Особую опасность представляют атаки смещения табличных данных, приводящие к искажению статистических свойств выборки, акустические атаки, при которых вредоносный триггер внедряется в аудиосигнал и остается незаметным для человека [1].

Смещение распределений возникает, когда статистические характеристики признаков в обучающей выборке отличаются от ожидаемых или целевых распределений. Такое смещение может быть как непреднамеренным, так и результатом злонамеренного вмешательства [2]. В связи с этим возникает необходимость формализации методики проверки данных, применимой к различным типам входной информации.

Предлагаемая методика предназначена для систематической проверки данных, используемых при обучении моделей машинного обучения, и направлена на выявление смещения распределений, подмены меток, отравления данных и внедрения акустических backdoor-триггеров []. Методика является модально-независимой и адаптируется к особенностям табличных и аудиоданных.

Этап 1. На первом этапе производится описание и формализация проверяемого набора данных.

Этап 2. На данном этапе выполняется проверка базовых статистических свойств данных. Для табличных данных - оценка распределений признаков, анализ выбросов и пропусков и проверка корреляционной структуры.

Для аудиоданных - анализ временных характеристик сигналов, оценка спектральной плотности и контроль диапазонов частот и уровней шума.

Этап 3. На третьем этапе проводится анализ смещения распределений признаков и целевой переменной.

Этап 4. Этап направлен на выявление нарушений в распределении целевой переменной и возможной подмены меток.

Этап 5. Обнаружение отравления данных (Data Poisoning).

Этап 6. Проверка аудиоданных на наличие акустических атак (этап применяется к аудиоданным и моделям, использующим акустический ввод).

Этап 7. На заключительном этапе оценивается устойчивость модели и достоверность результатов проверки.

В рамках исследования было разработано программное обеспечение, реализующее предложенную методику проверки данных. Программный комплекс предназначен для

автоматизации этапов анализа данных, выявления смещения, аномалий и признаков атак и формирования отчётов по результатам проверки.

Архитектура программного обеспечения модульная и включает модуль анализа табличных данных, анализа аудиоданных и визуализации и отчётности. Программное обеспечение поддерживает расширение и адаптацию под различные типы моделей и форматы данных.

Выводы

В данной работе представлена методика проверки данных, используемых в машинном обучении, ориентированная на выявление смещения распределений, подмены меток, отравления данных и акустических backdoor-атак. Экспериментальное исследование показало, что предложенный подход позволяет существенно повысить надежность моделей машинного обучения и снизить риск скрытых атак, направленных на данные. Разработанное программное обеспечение подтверждает возможность практической реализации методики и ее интеграции в процессы аудита и сопровождения систем искусственного интеллекта.

В дальнейшем планируется развитие методики в направлении автоматизации принятия решений, расширения набора поддерживаемых модальностей и формализации критериев безопасности для сертификации систем машинного обучения.

Литература

1. AI Secure Agentic Framework Essentials [Электронный ресурс]. – Режим доступа: https://yandex.cloud/ru/security/ai-safe (Дата обращения 10.02.2026).

2. ГОСТ 59276- 2020 (ИСО 787/1-82). Системы искусственного интеллекта. Способы обеспечения доверия. Общие положения. Введ. 23.12.2020. Стандартинформ, 2021. 16 с.

3. Менисов А. Б., Ломако А. Г., Дудкин А. С. Метод защиты нейронных сетей от компьютерных бэкдор-атак на основе идентификации триггеров закладок //Научно-технический вестник информационных технологий, механики и оптики. – 2022. – Т. 22. – №. 4. – С. 742-750.